SMU Data Science Review

Volume 3
Number 3 *Fall 2020*

Article 2

2020

Well Oiled Machine: Classifying Machinery Performance Reductions Using Work Order Data

Jacob Brionez
Southern Methodist University, jbrionez@smu.edu

Amber Burnett
Southern Methodist University, aburnett@smu.edu

Cho Kim
Southern Methodist University, crkim@smu.edu

Scott M. Whitney
scott.whitney@exxonmobil.com

Thomas N. Anderson
thomas.n.anderson@exxonmobil.com

See next page for additional authors

Well Oiled Machine: Classifying Machinery Performance Reductions Using Work Order Data

Authors

Jacob Brionez, Amber Burnett, Cho Kim, Scott M. Whitney, Thomas N. Anderson, and Sumeet Treehan

Well Oiled Machine: Classifying Machinery Performance Reductions Using Work Order Data

Jacob Brionez[1,2], Amber Burnett[1], Cho Kim[1]
Scott M. Whitney[2], Thomas Anderson[2], Sumeet Trehan[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] 22777 Springwoods Village Parkway,
Spring, TX 77389 USA
{jbrionez, aburnett, crkim}@smu.edu
{jacob.j.brionez, scott.whitney, thomas.n.anderson, sumeet.treehan}@exxonmobil.com

Abstract. Work Order (WO) data from System Applications and Products in Data Processing (SAP) software contains valuable information about what WOs intend to accomplish. Using SAP work order data, with time-series machinery sensor data combined into the same dataset, provides an opportunity to optimize prediction models to increase performance. Ideally, WO data can be utilized to help predict machinery's anticipated performance and can help prioritize a WO among others based on the anticipated machinery performance. It is possible to identify anomalies in pump sensor data using the Isolation Forest algorithm as the method for anomaly detection. The relationship between the sensor data and the WO data is not straightforward due to scheduled maintenance programs, causing anomalies in the data and periods where a pump has experienced higher than normal performance. Autoregressive integrated moving average (ARIMA) provides additional insight from a time series perspective but may not necessarily provide different results. However, some anomalies did show that some advance notice or other factors that could be used for elevating the priority of a work order. Further analysis in what is considered to be a "good" and a "bad" anomaly may need additional research to enable a more efficient approach to detection with respect to WO prioritization of "bad" anomaly data.

1. Introduction

System Applications and Products in Data Processing (SAP) software is the backbone of information for many large companies. It handles not only financial information but also logistical information that affects the daily operations of a large organization. This research focuses on the work order (WO) information entered into the SAP system by utilizing the details associated with equipment issues occurring at oil and gas production facilities. WOs have evolved from large paper-driven systems to massive digital databases organized in the form of relational data. WOs contain information regarding a task that needs to be performed and the labor, materials, tools, and services that are required for the task. Because of this traditional manual approach, illustrated in Figure 1, the digital system is still rooted in manual entry and management procedures

performed by various personnel ranging from operators to engineers to managers. This manual aspect makes the system prone to non-standard entry of data and a variety of entries prone to errors in interpretation. Over time, cleanup of these systems involved exercises in data standardization, duplicate reductions, and deletion of obsolete WOs (Hodkiewicz & Ho, 2016). This data cleaning is driven by the need to utilize data in these databases for a higher purpose. This research attempts to use SAP WO data for the higher purpose of increasing the productive hours and performance of the equipment focused on in this study.

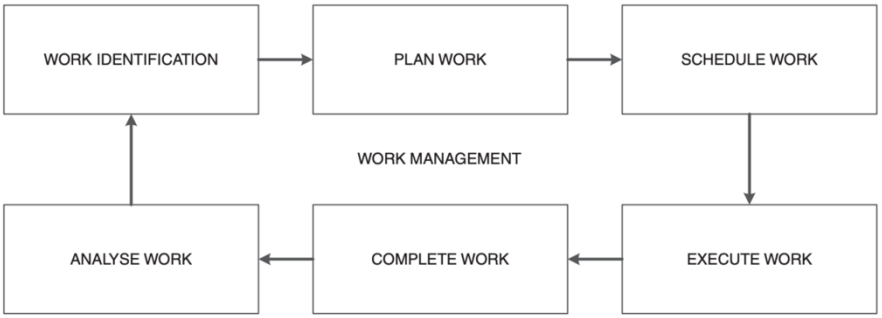

Fig. 1. Six Stages of Work Maintenance Management.

The prioritization of work order data can help increase system performance. Various industries, such as power plants, manufacturing, and oil and gas industries, require dependence on facilities' reliability and availability for them to be successful (Cline, et al., 2017). Since these various industries are large scale operations with many moving parts, it is crucial to prioritize WOs correctly to minimize the impact to the health and safety of employees as well as the environment but to also minimize the impact of operations. Within the energy industry, facilities must consider the sequence of work orders. Work orders completed in an unsystematic sequence can waste time and money for the company during unnecessary downtime in production.

In order to examine WO creation, this research focused on analyzing sensor data from a crude transfer pump within a separation process to see if anomalous sensor data could be used to predict the need for a WO. The magnitude of the anomaly from the sensor data may imply the severity of a problem within a piece of equipment. This information potentially could be used to automate creation of WOs for equipment. Depending on the magnitude of the anomaly, this information could be used to determine the severity of a potential problem and thus help prioritize WOs.

Strategically, prioritizing WO data contributes to saving companies millions of dollars by preventing the loss of revenue, preventing the mismanagement of maintenance programs, and ensuring the reliability of the company's reputation (Kohli, 2018). Currently, work orders are prioritized manually based on operations experience at the facility. This study will show how the association of work order data with machinery performance metrics will help to prioritize work orders more efficiently.

A work order is typically entered as a reaction to an already existing issue or as part of a maintenance program. They are rarely entered in anticipation of an issue. Advances in technology have enabled more profound insights that produce actionable results through the use of advancements in hardware and software. Utilizing centralized servers to collect and analyze data, we are now able to perform better forecasting of equipment failures to schedule maintenance before they occur. As shown in Figure 2, the North American Energy and Utilities market alone is projected to be $251.3M USD by 2022 (Otto, 2019).

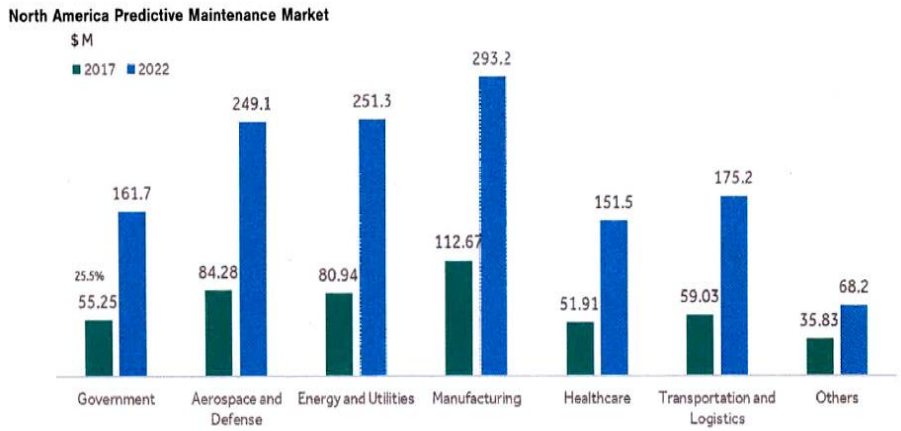

Fig. 2. North America Predictive Maintenance Market. This figure shows the major predictive maintenance market projective increase over 5 years.

As a result, goals to reduce lifecycle costs for operations and maintenance are more and more realistic to the point that asset optimization increasingly affects corporations' financial calculations (Dandashly, 2012).

The usual suspects for performance degradation are time-dependent, commonly due to wear and tear or slow deterioration of materials over time. Others are perhaps not as slow but are still time-dependent, for example, deterioration due to lack of lube oil or particles in lube oil, which wears parts more quickly. With the WO data, the research can now take into account issues that may occur specific to certain models or equipment running under similar application to predict performance degradation on a more granular basis and provide more specific action items to appropriate field personnel for more efficient remedies reducing error, downtime, and risk.

Prior research is based on collecting information on mechanical failures. The intention is to use an improved form of measure such as performance for the research. El-Abbasy et al. (2014) used deterioration to gauge the effect of each factor on the overall condition being monitored. Using the metric of performance deterioration to prioritize the order of work orders helps to comprehend the severity of failure better. Since failures are not always a complete shutdown of machinery, it is better to use its percentage reduction or deterioration as a metric.

2. Related Work

Researchers have pursued different methods regarding machinery degradation and prioritizing WOs. None of the prior studies bridged machinery degradation and how it relates to WOs. This current study helps bridge the gap between detecting machinery degradation and how it relates to WOs.

These related works are organized by each section below according to their contributions to the study. Literature in each section involves scholarly studies regarding areas of Maintenance Management, Data Cleaning and Mining Rules-Based Approach to WOs, Maintenance Request Prediction, Machine Learning Concepts for Equipment Life Monitoring, Machine Learning Methods, and Anomaly Detection in Time Series.

2.1 Previous Maintenance Management

Maintenance management is a critical and essential practice done by many companies as a way to reduce equipment downtime and labor. Yang et al. (2007) used a quantitative approach to measure performance by summing time spent at each workstation in an assembly line. For their research, they applied a "no shut off" rule meaning that the product continued down the assembly line even though a part of the line failed until the end of the line is reached or if the equipment failure causes the line to prematurely halt. Yang et al. (2007) defined system value as a "summation of all part values existing in the system at a given moment." Below is a mathematical expression of system value where the variable, i, represents the station, v_i represents the part value, and $C_i(t)$ represents the number of parts in station i, equation 1.

$$W(t) = \sum_{i=1}^{n} v_i C_i(t) \quad \textbf{(1)}$$

System value and production effort are directly correlated, therefore "the higher the system value is, the more production effort has been done" (Yang et al., 2007). Using their system value-based approach, they argued that prioritizing WOs based on system value gave maximum productivity.

The importance of maintenance management was also emphasized by Zemenkova et al. (2016) in their research. Since an oil and gas pipeline system is large and complex, a lot of considerations need to be made to operate safely and reliably. Their objective was to improve their decision-making practices related to preventative maintenance. The "generally accepted standards" for equipment maintenance has been to follow repair schedules, but Zemenkova et al. (2016) argued that monitoring machinery helps prolong equipment service life. Advances in machinery have made it possible to perform real-time analytics on condition monitoring equipment efficiently. Their research examined the following parameters for a gas compressor unit: "effective power, fuel consumption per hours, the rotational speed of rotors, temperature characteristics (for example, temperature of gases before the turbine, input and output values of gas temperature and pressure), lube oil pressure and temperature, gas

composition" (Zemenkova et al., 2016). Their theory was based on "random process overshoot" to determine if the equipment exceeded the permissible level during operation. They used a smoothed average that is calculated by "creating series of averages of different subsets of the full data set" (Rinfret, 2019). Using this type of forecasting helps to justify recommendations for downtime repairs and proper planning for material and man hours (Zemenkova et al., 2016).

2.2 Data Cleaning and Mining Rules-Based Approach to Work Orders

Numerous quality and reliability issues surround work order data from maintenance computerized management systems. The quality issues arise within the data, which is mined and cleaned poorly, resulting in inaccurate results and detection, which comprise the accuracy of maintenance facilities. Practicing identifying and exploring data quality issues was a key area addressed by Hodkiewicz and Ho (2016). The researchers were able to see that among 85,073 WO records across three organizations, up to 37.2 percent of these were assigned incorrect WO entries (Hodkiewicz & Ho, 2016). They concluded that inaccuracies in WO can be from intentional and unintentional causes. To help correct these problems, the researchers studied the life cycle of a maintenance order in great detail and realized where discrepancies, inaccuracies, and mistakes were broken within the system as seen in figure 3.

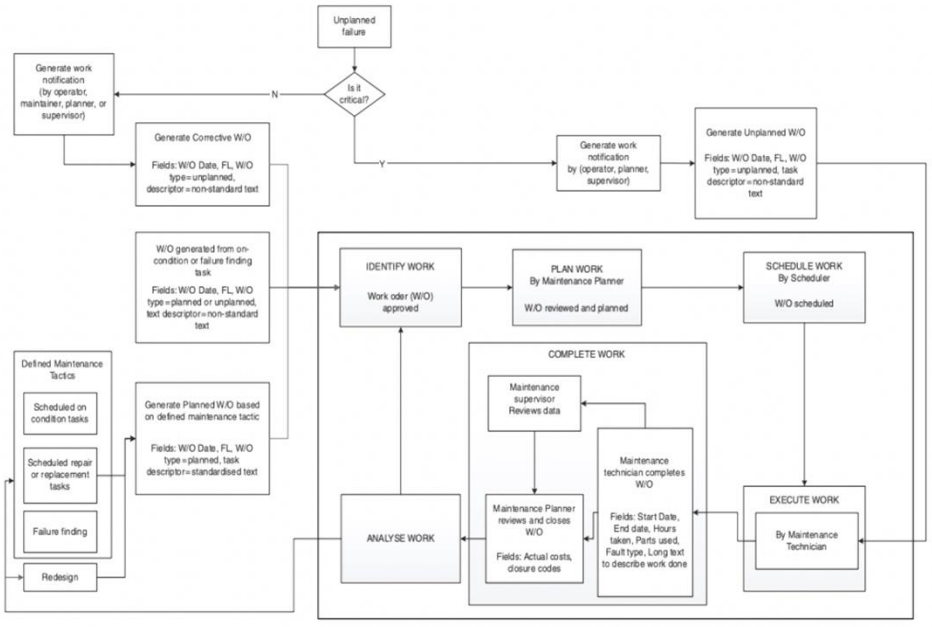

Fig. 3. Life Cycle of Maintenance Work Order.

Their discoveries helped them to successfully develop a rules-based approach of rule libraries that helped in cleaning and mining the data along with effective maintenance practices which involved identifying and fine-tuning organizational wide practices to follow these rules. These rules and maintenance practices were successfully implemented, analyzed, and monitored with modification or addition of rules organization wide to help with failure analysis at facilities. The rules-based system was syntactic versus semantic. The syntactic rules-based system tool was developed, called DEST, and is preferred to semantic based rules as it eases the implementation, mapping, parsing, and rules execution.

2.3 Predicting Maintenance Requests

With the widespread use of analytics and machine learning, many researchers proposed a proactive approach to maintenance by predicting equipment failures. Predictive maintenance is defined as "maintenance performed based on an estimate of the health status of a piece of equipment" (Susto et al., 2015). Using ThingWorx API, Parametric Technology Corporation was able to integrate inspection data obtained from their ThingWorx API to identify potential failures in connectors for oil and gas equipment. The maintenance inspection data Parametric Technology Corporation used came from "two families of oil and gas equipment, swivels and valves, for two downstream customers" (Cline et al., 2017). They have over 123 different swivels and valves and 206 different models of equipment that are tracked from the beginning of their service contract until the equipment is scrapped or the service contract expiration is reached. Their artificial neural network with three layers performed the best out of the other machine learning models with a 46% captured failure rate, which was better than their existing model, which was based on employee reports, which had a 2% capture failure rate. They incorporated their model into a web application to create a visualization tool that helped classify assets by risk for end-users and incorporated map data to "[enable] new service delivery models that may be location and risk driven" (Cline et al., 2017).

Another approach to predicting equipment failure was made by Kohli. Kohli (2018) integrated historical data and SAP process data in machine learning models to predict equipment failure. The author examined "274 instances for 39 equipment and each instance was associated with 11 features." For pumps, Kohli operated under the assumption that the lifespan of a pump is 15 years. Using classification, Kohli categorized each instance as preventative, corrective, or breakdown. Kohli used a 75%/25% train-test split where n = 52 using the Weka resampling filter with no replacement. The model that performed the best from the research, with 98% accuracy, was a combination of Decision Trees (DT) and Support Vector Machines (SVM) using 10-fold cross-validation (Kohli, 2018). As part of the 10-fold cross-validation, Kohli used boosting, bagging, and stacking techniques to negate bias and variance in the research. Since this study will be using WO data obtained from SAP, Kohli's research showed that SAP data, along with equipment reliability data, could be used for machine learning to classify and predict equipment reliability. We hope to expand on Kohli's research and use advanced machine learning techniques to classify WOs and predict performance. Machine learning algorithms can be utilized on big sensor data to provide predictions for continuous and categorical response variables (Kejela et al., 2014).

2.4 Machine Learning for Equipment Life Monitoring

In the age of big data and the Internet of Things (IoT), there are many sources of data collected in all aspects of the oil and gas industry ranging from sensor data to image data. Mohammadpoor and Torabi found that big data analytics were being used across different stages of oil and gas production and found that it has been used to "[optimize] the performance of electric submersible pumps" (Mohammadpoor & Torabi, 2019). It is crucial to use the correct associated data for that machinery to take into account the effect of systems directly connected to it (Almasi, 2012). For machines to continue operating effectively, condition monitoring is an important aspect to consider for performance. General Electric uses artificial intelligence and machine learning for tool life modeling in machining (Aggour et al., 2019). Deterioration is a vital feature to consider in our modeling because it can have a direct impact on performance.

Tan et al. (2016) pointed out in their research that despite the "vast amounts of data [that] have been collected" regarding fatalities and injuries in the upstream oil & gas industry, those previous researchers were unable to leverage the data to find high-level trends due to the fragmented nature of the data. They also pointed out that oil and gas industry data can contain 20 petabytes of data, and "many operational aspects of the oil and gas industry are also generating significantly more data than they used to" (Tan et al., 2016). With the evolution of modern technologies and tools available for big data today, there are many opportunities for this study to use newer analytical techniques to discover trends.

2.5 Machine Learning Methods

Orrù et al. (2020) developed a supervised machine learning model for "fault diagnosis of rotating machinery in the oil and gas industry." They used data collected from a centrifugal pump in the production line at the SARLUX refinery located in Sarroch, Italy. Their dataset spanned over five years and consisted of specifically "centrifugal pump operations-related sensor readings" for eight different sensors for flow rate, bearings vibration, axial displacement, and motor coil temperature. Since their analysis was done on the Konstanz Information Miner (KNIME) which is an open-source analytics platform, they were able to leverage the Missing Values node in KNIME which filled in the missing values using a linear interpolation technique. Orrù et al. chose this method because it was "simple and effective for time series data." To account for downtime periods, Orrù et al. filtered out the data during downtime periods and data related to start-up periods. They defined downtime as times in which "machine has been shut down due to maintenance operations or other issues."

In order to avoid overfitting their model, Orrù et al. (2020) employed hyperparameter optimization techniques like grid search and k-fold cross validation. For their case study, they determined that three-fold cross validation "ensured a good representation of the input dataset" since failure events they were trying to classify was not represented evenly in their data. Orrù et al. (2020) compared two different machine learning classification techniques: Support Vector Machines (SVM) and Multiplayer Perceptron (MLP). Their SVM model performed with 98.1% overall accuracy and their MLP artificial neural network (ANN) model performed with 98.2% overall accuracy.

Liu et al. (2018) confirmed that both SVM and artificial neural network techniques performed well related to fault diagnosis for rotating machinery. According to Liu et al. (2018), "SVM has excellent performance in generalization, also with few training data" and "data from two or more categories can always be separated by a hyperplane" which contributes to producing a high accuracy for fault diagnosis. According to Liu et al. (2018), ANN "mimics the human brain structure" and due to its versatile nature, it can produce "good fault diagnosis performance in many rotating machinery applications."

2.6 Anomaly Detection in Time Series Methods

Anomalies or outliers are detected continuously, whether it be in everyday life or in a statistics class. They are considered to be things, or data points, that deviate from the norm or the normal distribution of data. Krishnan's research focused on anomaly detection with time series forecasting problems solved by algorithms like Seasonal Auto Regressive Integrated Moving Average (SARIMA), Long Short-Term Memory (LSTM) recurrent neural networks, and Holt-Winters method for triple exponential smoothing. The research focused on the estimation of future needs with contemporary data (Krishnan, 2019). Krishnan learned through his findings forecasting could help detect anomalies. Krishnan studied the SARIMA and LSTM methods for anomaly detection with time series forecasting. The SARIMA algorithm predicted that the actuals were on most ordinary days and captured the trend from the spikes in high accuracy. However, the SARIMA algorithm would be tedious and exhaustive in execution and time. The LSTM method involving the recurrent neural network worked well for the metrics. However, an auto ARIMA was used to help with forecasting technique problems where the algorithm is unable to point out the actuals. This is a problem that they identified in which every metric needs to be fine-tuned with parameters so that the anomalies are detected precisely while using anomaly detection forecasting.

Isolation forest is an additional anomaly detection algorithm that assigns an anomaly score to each sample in a dataset. The algorithm can isolate observations by randomly selecting a feature and selecting a random split value between the minimum and maximum values of the feature. The number of splits is equivalent to the path length from the root to terminating nodes, and path length is the measure of normality and the decision function. Path lengths that are shorter for a particular sample are highly likely to be anomalies. A significant benefit of using isolation forests is that it requires low memory and is ideal for large datasets. Figure 4 below shows an example of an isolation forest model's output from Sci-Kit Learn (Pedregosa et al., 2011).

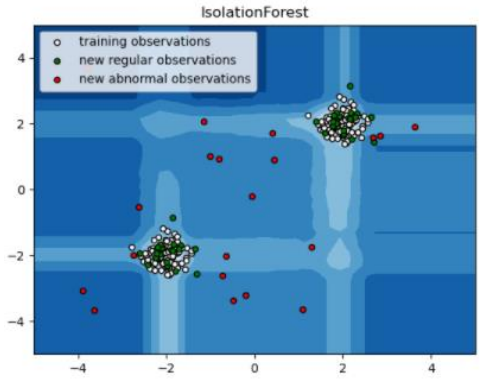

Fig. 4. Isolation Forest Plot Example

In another study, Chen and Overstreet (2019) developed a real-time anomaly detection for time series at Pinterest. Chen and Overstreet's vigorous and expandable anomaly detection system helps engineers interpret and react to problems as they occur and does not interrupt the Pinterest business or users (Chen & Overstreet, 2019). Chen and Overstreet updated their Stats board dashboard in-house metrics and dashboard system through brute force, scheduled, event-driven, and online updates. The dashboard upgrade research highlighted requirements for anomaly detection observability, which is very insightful and necessary when building a robust system. They found that minimizing false positives is possible by only alerting on the most severe anomalies. Also, time-sensitive incidents must manifest as high priority so that they are actioned as soon as possible. A system must also be robust enough to support data scaling for millions of data points needed for anomaly detection and missing data scenarios. The use of an action item approach to addressing what is most important will help this study succeed in anomaly detection with time series.

3. Data

WO data is continuously generated and stored in an Enterprise SAP system locally for all global facilities. Major data restructuring has moved this data to a Data Lake for analytics consumption. The time-series data is pulled from machinery sensors locally at each of the facilities and stored in an OSISoft PI Historian (Fig. 5). The OSISoft PI historian data is viewed and exported using a process manufacturing software called Seeq. Seeq is an application that can pull data from multiple data sources and clean time series data very quickly and easily (Talmadge, 2017). The Seeq Workbench provides data visualization, modeling, and analytics tools. This research focused on a single crude transfer pump within a separation process. The data was exported from Seeq between September 30, 2018 and September 29, 2020. Sensor readings for the pump were taken in different time intervals ranging from 10 seconds to 45 seconds per interval. The dataset contained over four million records for nine features including the

pump's on/off status and date time stamp of the sensor reading. The exported data from Seeq has a timestamp for each row of data and the sensor readings for volume, power, temperature, pressure, flow, output, and status as the column name. The SAP data was derived using NLP algorithms on the SAP WO database to clean and categorize the WOs. This activity was performed prior to this research, and the results are used to relate the SAP data to the sensor data using the location of the facilities and the equipment names specific to those facilities. Once this relationship was established, all other attributes could be related to the sensor data for analysis.

Early exploratory data analysis revealed that SAP WO timestamps do not necessarily correlate with the adverse activity in the sensor data. Additionally, some of the changes in data during the WO start and end dates are not inclusive of the actual event data that prompted the entry of the WO itself.

Utilizing the Snowflake SaaS product, this research will consume the data in relational databases to curate the data for use in visualization or further analysis. This study used Python for the analysis. The transformation of the data may be necessary to evaluate the performance degradation of the machinery during the specified periods. These will be independent variables in this study. Specific information from WOs and other associated sensor data will be the dependent variables in this analysis. The Time Series elements are considered in three parts: the period before the entry of the WO, the time period between the WO entry and close dates, and the period after the WO Close date for baseline data, degradation data, and anticipated performance increase data, respectively.

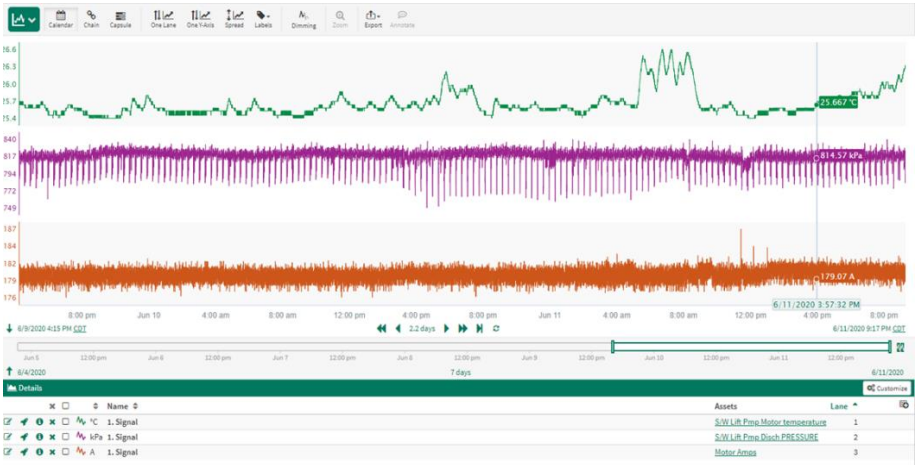

Fig. 5. Time Series Pump Data within the Seeq Dashboard.

Figure 5 depicts temperature in Celsius, pressure in kPa, and current in amps. The final form of the data contains the sensor information for the equipment as it relates to the WO data obtained from the SAP database. Each timestamp will be associated with a sensor value as well as attributes from SAP such as equipment type, manufacturer, model, WO status', Severity, Issue description, and method of resolution. The data will

then be filtered to show only the specific equipment and issues that we will use for our data analysis. These filtered data sets are then exported as .csv files to be analyzed using Python/R.

4. Methods

Building a comprehensive data lake for large companies is a significant undertaking and a continuous effort. The data used for this research is from two databases that will eventually coexist in the corporate data lake. Since the WO data and the sensor data do not currently coexist in the data lake, considerable effort was put into obtaining and correlating the datasets, not to mention legal discussions to request usage of the data for this research. The global dataset contained data for all machinery located at all facilities. SAP data for WO's that had common descriptions for issues for machinery that had common machinery types and manufacturers helped to determine which data to analyze. Pumps were chosen due to the high number of standard units in facilities and common failure modes. It was necessary to review the process flow diagrams relating to the equipment to determine if sensor data was common for different machinery in different locations.

4.1 Data Processing

The data used for the research contained over 4.1 million records spanning two years and contained sensor readings from a pump within a crude oil separation process. The most amount of null values for any feature was under 40,000. Since the number of null values was less than 1% of our total dataset, the missing data was removed for the analysis. Figure 6 below is a visual representation of the dataset. The horizontal white lines represent missing data. The areas of solid gray outweigh the missing data which further validates the decision to remove missing data.

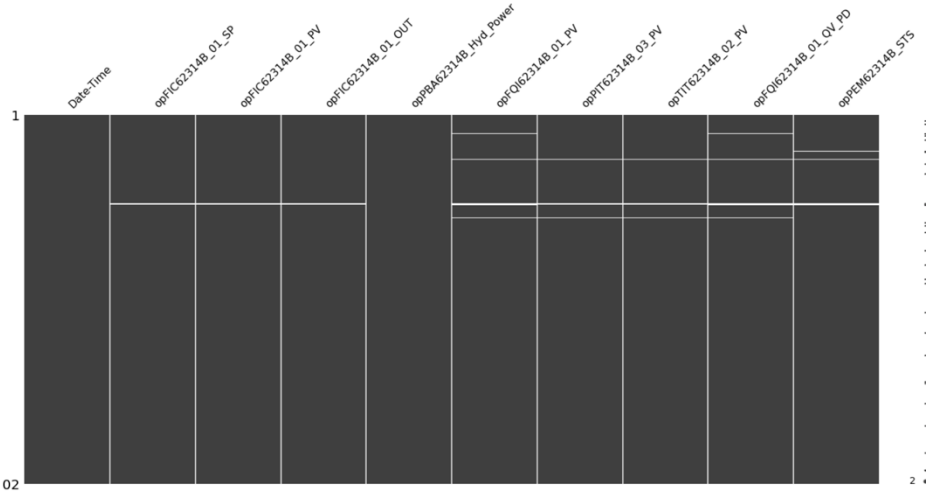

Fig. 6 Visualization of Missing Data in Pump Dataset.

In order to use the dataset in the machine learning models, the sensor data where the pump was off needed to be excluded from the analysis. There were 2.45 million records remaining for analysis after excluding the data where pump status was off. In the initial approach of the research, the focus was centered around time periods before and after WOs. Anomaly detection may help to correlate the prior data that was not taken into account. Rather than begin with the WO start and end date windows, locating the anomalies for the time series of the equipment and back associate the event time frame to see where the event data overlaps the WO start and end dates. This procedure helps to classify the events but not necessarily the severity of the event/WO. To gauge the severity of impact for the event, observing the production/performance reduction data for the equipment would have to be done using the dataset from the event time frame.

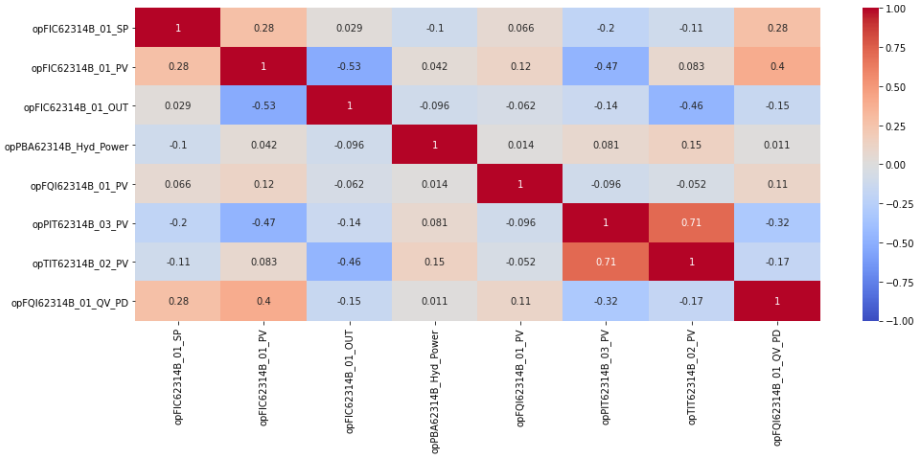

Fig. 7. Pearson's r Correlations Heatmap. This figure describes Pearson's r Correlations measuring the strength of the association between variables.

Figure 7 describes the Pearson's r correlations ranging from a positive correlation (1.00) to a negative correlation (-1.00). This figure is a visual representation of all features within the dataset that summarizes correlations for each variable. The heatmap concludes some variables are highly correlated within the data such as opTIT62314B_02_PV is highly correlated with opPIT62313B_03_PV with a correlation score of 0.71. This is a strong warning indicator for our data analysis. Multicollinearity may be an issue here and impact our results when trying to fit the model (Sanchez Lafuente, 2020).

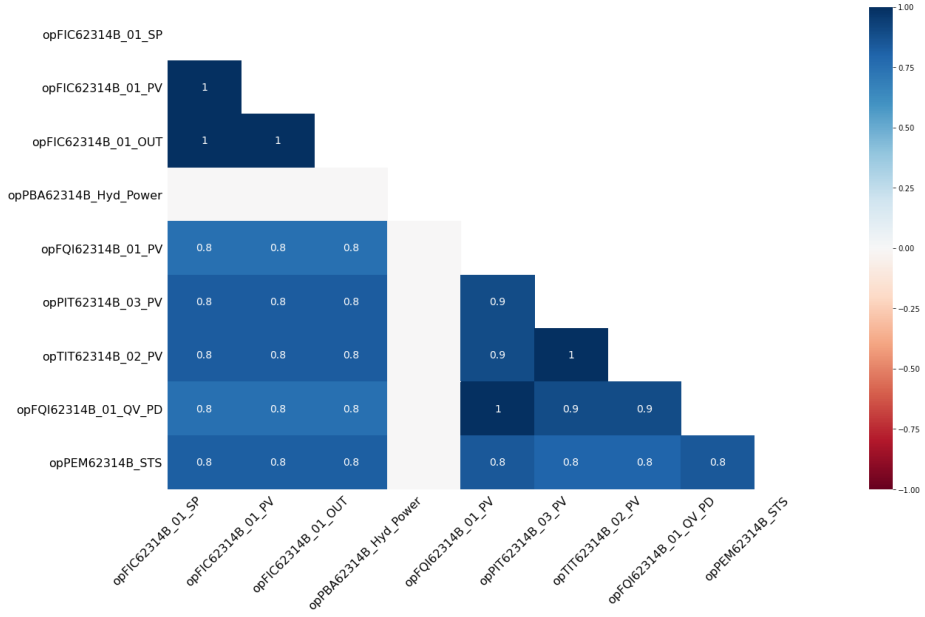

Fig. 8. Missing Values Heatmap.

Figure 8 shows the missing number correlation heatmap which visualizes "how strongly the presence or absence of one variable affects the presence of another" (Bilogur, 2018). The correlation can range from -1 to 1. A correlation of -1 means for this particular set of features if there is a missing value, the other feature will not have a missing value. A correlation of 0 means that there's no effect if data is missing or not between the two features. A correlation of 1 means that if there is data for one feature, there is data for the other feature and if a value is missing from one feature, it's missing from the other.

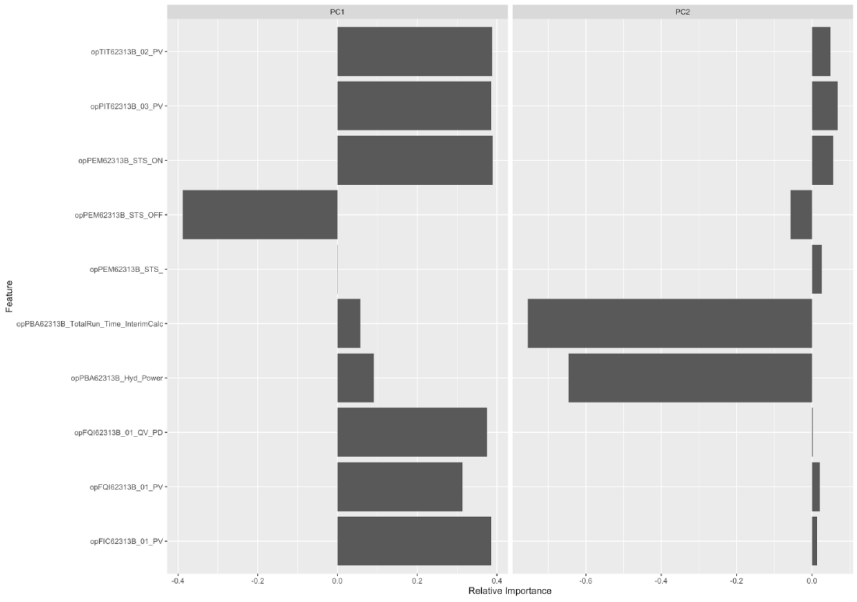

Fig. 9. Principal Component Analysis (PCA) Feature Importance Figure. This figure displays the PCA feature importance.

Figure 9 displays the relative importance of specific features. Within the Principal Components there was a 65% cumulatively explained variance for PC1 and a 76% cumulatively explained variance for PC2. The variables with the greatest interest on separation of samples is shown for PC1 and PC2. The PC1 is less desirable to consider seeming most features are favored to the positive side. The PC2 is influenced significantly on the negative side by TotalRun_time and Hyd_Power features. We will further investigate these features and their significance with other methods.

4.2 Method for Analysis

The Isolation Forest algorithm from Scikit-learn was utilized for analyzing sensor data on a univariate basis for anomaly detection. This was able to provide ranges in the sampled data that were anomalies based on anomaly scores assigned to each of the anomalous data points. The anomaly ranges derived were then visualized in Seeq with their associated time series plots. Using these visualizations it was possible to relate the anomalies to the documented events in the SAP WO system.

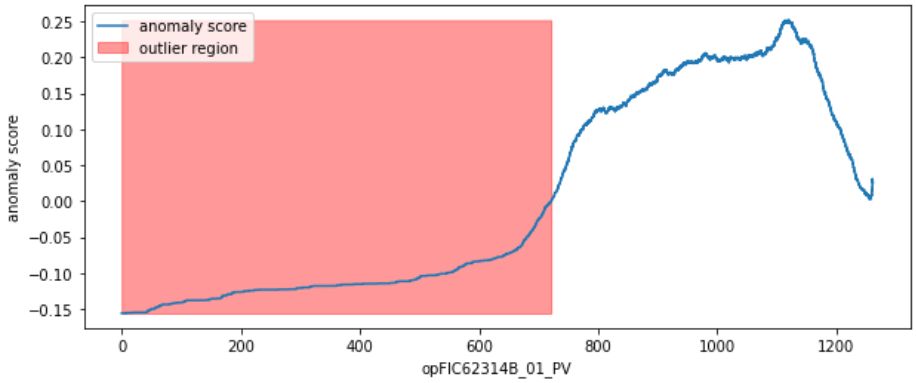

Fig. 10. Visualization of the Anomaly Score from Isolation Forest

Figure 10 above is for the opFIC62314B_01_PV feature, which is one of the flow readings measured in m^3/h, for the crude transfer pump. The red area in figure 8 represents the range of values for the opFIC62314B_01_PV feature that was considered an anomaly. This red range was entered in Seeq to visualize where the anomalies occurred. This process was repeated for all nine features.

5. Results

This current study leveraged different anomaly detection algorithms such as Isolation Forest, ARIMA, and VAR. Isolation Forest produced more applicable results than ARIMA and VAR due to the number of variables being highly correlated. The nature of the time-series data did not lend itself to ARIMA and VAR well because of erratic characteristics.

The use of sensor data to do more than trigger alerts for boundary conditions can be done utilizing contextual information from other data sources. For this research, the context was provided using WO data. Data from actioned items for issues that have occurred and were remedied were used as attributes for the sensor data.

For the Isolation Forest model, the tuning parameters were changed from the defaults for optimal results. Using the default parameters for the model resulted in a noisy graph that had multiple outlier regions. The output using the default parameters for the same variable as figure 10 is below. Figure 10 represents the fine-tuned model for this study and figure 11 depicts the initial run. The area defined as the normal range in the initial model is much smaller than the area in the final model. Using a brute force approach, the dates from the anomaly ranges in the initial model was compared with the WO data on or around those dates. There were many instances from the initial model that there was not WO data for therefore we determined it was not a valid anomalous range and fine-tuned the model. The best performing Isolation Forest model used 300 isolation trees and a contamination value of .01.

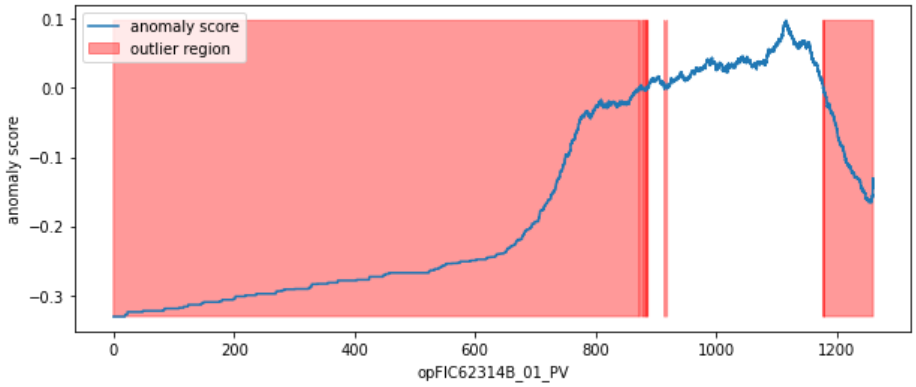

Fig. 11. Visualization of the Anomaly Score of opFIC62314B_01_PV for Default Parameters

Utilizing the Isolation Forest algorithm to identify anomalies for each of the features, it was possible to visualize them. Using these trends, this current study located anomaly range values that were then visualized in Seeq further. Many of the anomalies coincided with the scheduled maintenance programs for the equipment, verified with the use of the SAP WO data. Some anomalies coincided with the increased performance after some of the planned maintenance tasks were performed. This showed that the problem occurred before the WO was created. It was assessed that these anomalies were related to the WO by the timestamps recorded for Order Entry Date, Order Start Date, and Order Completion Date.

For example, low temperature anomalies preceded an installation of insulation WO that was entered February 15, 2019. The WO was completed on February 24th. The pump was shut down during this time due to low flow. After the insulation was installed, the temperature was stabilized and no anomalies were detected for two months. The figure below shows that there were no anomalies, shown as green bars at the top of the trend after the insulation was installed. Furthermore, as a baseline the research utilized mean + or - two standard deviations shown as yellow bars. The Isolation Forest model was more sensitive to anomalies than the baseline anomaly range.

Fig. 12. Visualization of bearing temperature in Seeq with anomalies

6. Discussion

To predict the need for a WO and prioritize it in a way that was not subjective but done prescriptively, known anomalies were taken and assigned an identity (derived from the associated WO description) that could be used as training data for prediction. Because WO data by itself did not relay the impact it had on the health of the facility, there also needed to have been a way to assign values to each WO created based on the equipment's effect on the facility.

The results indicated the anomalies found for the purpose of classifying events types for prediction could be utilized. In the future, further research could be done to automate a workflow where detected anomalies associate with a WO type and record an event time frame for use in building training sets for prediction/earlier detection of similar issues in real time situations.

Due to a high correlation between pressure and flow, the anomaly ranges from the model detected abnormally high amounts of anomalies compared to temperature as shown in figure 13 below. This was due to the high variability in the pressure and flow data. Over the time period of data the research analyzed, there were many false positive anomalies from the Isolation Forest model and the baseline model. This in part is due to the amount of manual changes in the flow setpoint for the pump. Even though the setpoint was set manually many times, temperature was resilient in producing insightful anomalies that could be used for prioritizing WOs in the future. Pumps in other applications may not be so affected by setpoints enabling the use of additional parameters for anomaly detection.

Fig. 13. Visualization of pump discharge pressure with anomalies in Seeq

7. Ethics

The research team took ethical considerations into account at every stage before, during, and after the project. Currently, there is no governing set of rules overseeing the actions of Data Scientists in the United States, so it was necessary to seek out another form of guidance in order to ensure that all research practices used in this project were ethical.

Saltz & Dewar (2019) identified ten important ethical questions that data scientists should ask. The research team took care to address each of these ethical questions in the design and implementation of the project. Ethical accountability was continually achieved by making sure the model was disproportionate, and that it accurately presented the results. The data was obtained appropriately, and there were no privacy issues concerning the data. The data was collected through a reputable source, was clean, and also understandable. Considering the data has no one's personal information there were no concerns with anonymity. The data was continually monitored to ensure no harm came to human life.

Within the oil and gas industry specifically, ethical decision making has been implemented as an optimization process with things such as human safety of workers a priority over corporate profits (Carpenter, 2018). This warranted in the research such things as the minimization and mitigation of harm that aligns with the efforts of the oil industry's current standards.

It was important to take in account various relatable laws and regulations that applied to the particular area of study. The data used in this project does not require such scrutiny since the industry already goes above and beyond to ensure that existing laws and requirements are followed. For example, OSHA inspectors and company-

wide internal inspectors already ensured the compliance and fulfillment of expectations for the oil and gas industry. As a result, this research has no compliance concerns.

The research team took specific steps to maintain the integrity, reliability, and privacy of the data. The company's privacy is also maintained through the linkage of any data through trusted platforms. One of these was Seeq which ensured data quality vs data quantity. The Seeq tool provided a way for industries to speed up data cleaning with the immense amounts of data before populating their models. This afforded quality data with a reliable method to do so. The tool offered a more useful way to increase value while also improving operational productivity (Romanow, 2018). Within the data science decisions used during the research the integrity, reliability, and privacy of the data was maintained.

The security of the data was valid for its intended use within the project. It was knowledgeable of who owns the data, their rights and expectations of the data, and the scope of their intended use for the data. The research has been reviewed by the company and approved. The data is not compromised because of proprietary concerns.

To ensure the data was valid for its intended use, the entity approved it as suitable. The data validity was attained through data accuracy. One method used was within the research, was removing missing values with a valued statistical approach on the data which eliminated bias and took into consideration the fitness of the data's purpose and guaranteed data accuracy.

There are no sociological types of bias within the data. Throughout the research all decisions were attentively discussed, reviewed, and practiced extreme caution in decision making which ensured there were no potential modeler bias within the model building process. The project's decisions were looked at subjectively and various weighed options of optimizing values, which algorithm to use, model selection, anomaly detection, etc. and considered biases and prejudices that may be present during decisions that affected this study's final results. Effectively, the biases and prejudices were able to be deleted within the research.

The research team takes great pride in the level of transparency achieved in the data, the modeling process, and the reporting of results. The research team considered ethics the upmost important integration as data scientists, at every single step in the research process. The research guarantees that the analytical and subjective decisions that were made during our data project reflects the scale, accuracy, and precision of the data that was used in creating the model. With ethical considerations the research was able to systemize, defend, and recommend concepts of proper ethical conduct in relation to data and practices for the oil and gas industry.

8. Conclusion

Anomaly detection using time series and non-time series algorithms made it possible to isolate anomalies that could produce metrics to help in prioritizing work orders. Using performance as a metric to establish the effect of an actioned WO on the systems performance or level of production yielded mixed results. Overall, it was determined that planned maintenance is an effective way to maintain performance of pumps. However, most effective WO's had to do with lube replenishment or motor replacement

which accounted for the majority of workorders in SAP. It remains to be seen whether planned/scheduled maintenance could be modified based on this information since the idea of performing maintenance early is counterintuitive in most cases. However, it may be worth the effort to make a change to maintenance programs if the positive effect on production is beneficial enough.

9. References

A. Mohamed, M. S. Hamdi, & S. Tahar. (2015). *A machine learning approach for big data in oil and gas pipelines*10.1109/FiCloud.2015.54

Abbasi, T., Lim, K. H., & Yam, K. S. (2019). *Predictive maintenance of oil and gas equipment using recurrent neural network*10.1088/1757-899X/495/1/012067

Aggour, K., Gupta, V., Ruscitto, D., Ajdelsztajn, L., Bian, X., Brosnan, K., Natarajan, C., Dheeradhada, V., Hanlon, T., Iyer, N., Karandikar, J., Li, P., Moitra, A., Reimann, J., Robinson, D., Santamaria-Pang, A., Chen, S., Soare, M., Sun, C., Suzuki, Akane., Vinciquerra, J. (2019). Artificial intelligence/machine learning in manufacturing and inspection: A GE perspective. *MRS Bulletin, 44*(7), 545-558. 10.1557/mrs.2019.157

Almasi, A. (2011). Power plant condition monitoring. *Power Engineering, 115*(8), 60-63.

Almasi, A. (2012). Condition monitoring for rotating machinery: This valuable insight into the performance of pumps and compressors will help improve operation.(engineering practice). *Chemical Engineering, 119*(3), 55.

Bilogur, A. (2018). Missingno: A missing data visualization suite. *Journal of Open Source Software, 3*(22), 547. 10.21105/joss.00547

Carpenter, C. (2018). *Establishing data ethics in oil and gas operations.* Oil and Gas Facilities. https://pubs.spe.org/en/ogf/ogf-article-detail/?art=4780

Cerlioni, M.Anomaly detection in multivariate time series with VAR system monitoring in presence of serial correlation. *Towards Data Science,* https://towardsdatascience.com/anomaly-detection-in-multivariate-time-series-with-var-2130f276e5e9

Chen, K., & Overstreet, B. (2019). *Building a real-time anomaly detection system for time series at pinterest.* Pinterest Engineering Blog. https://medium.com/pinterest-engineering/building-a-real-time-anomaly-detection-system-for-time-series-at-pinterest-a833e6856ddd

Cline, B., Niculescu, R. S., Huffman, D., & Deckel, B. (2017). *Predictive maintenance applications for machine learning*. IEEE. 10.1109/RAM.2017.7889679

Dandashly, H. (2012). *Better data, better value operations.* Petroleum Economist. https://www.petroleum-economist.com/articles/upstream/technology/2013/better-data-better-value-operations-says-ge-oil-gas

El-Abbasy, M. S., Senouci, A., Zayed, T., Mirahadi, F., & Parvizsedghy, L. (2014). Artificial neural network models for predicting condition of offshore oil and gas pipelines. *Automation in Construction, 45*, 50-65. 10.1016/j.autcon.2014.05.003

Hodkiewicz, M., & Ho, M. T. (2016). Cleaning historical maintenance work order data for reliability analysis. *Journal of Quality in Maintenance Engineering, 22*(2), 146-163. 10.1108/JQME-04-2015-0013

K, D. (2020). *Anomaly detection using isolation forest in python.* Paperspace Blog. https://blog.paperspace.com/anomaly-detection-isolation-forest/

Kejela, G., Esteves, R. M., & Rong, C. (2014). *Predictive analytics of sensor data using distributed machine learning techniques*. IEEE. 10.1109/CloudCom.2014.44

Kohli, M. (2018). Predicting equipment failure on SAP ERP application using machine learning algorithms. *International Journal of Engineering & Technology, 7*(2.28), 306. 10.14419/ijet.v7i2.28.12951

Krishnan, A. (2019). *Anomaly detection with time series forecasting.* Towards Data Science. https://towardsdatascience.com/anomaly-detection-with-time-series-forecasting-c34c6d04b24a

Li, S. (2019). *Anomaly detection for dummies unsupervised anomaly detection for univariate & multivariate data.* Towards Data Science. https://towardsdatascience.com/anomaly-detection-for-dummies-15f148e559c1

Liu, R., Yang, B., Zio, E., & Chen, X. (2018). Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing, 108*, 33-47. 10.1016/j.ymssp.2018.02.016

Mohammadpoor, M., & Torabi, F. (2019). Big data analytics in oil and gas industry: An emerging trend. *Petroleum,* 10.1016/j.petlm.2018.11.001

Orrù, P. F., Zoccheddu, A., Sassu, L., Mattia, C., Cozza, R., & Arena, S. (2020). Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry. *Sustainability (Basel, Switzerland), 12*(11), 4776. 10.3390/su12114776

Otto, S. (2019). Predictive maintenance's role in improving oil & gas efficiency. *Pipeline & Gas Journal, 246*(2), 50.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825-2830.

Prabhakaran, S. *Vector autoregression (VAR) – comprehensive guide with examples in python*. Machine Learning Plus https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/

Rinfret, J. P. (2019). *Moving averages: A powerful tool in data science.* Medium. https://medium.com/@jprinfret/moving-averages-a-powerful-tool-in-data-science-67e3b25ce5db

Romanow, S. (2018). *How oil & gas operators are investing for innovation.* Automation.com. https://www.automation.com/en-us/articles/2018/how-oil-gas-operators-are-investing-for-innovation

Saltz, J. S., & Dewar, N. (2019). Data science ethical considerations: A systematic literature review and proposed project framework. *Ethics and Information Technology, 21*(3), 197-208. 10.1007/s10676-019-09502-5

Sanchez Lafuente, A. (2020). *Exploratory data analysis with pandas profiling.* Towards Data Science. https://towardsdatascience.com/exploratory-data-analysis-with-pandas-profiling-de3aae2ddff3

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics, 11*(3), 812-820. 10.1109/tii.2014.2349359

Talmadge, M. (2017). *Prediction.* Seeq Knowledge Base. https://seeq.atlassian.net/wiki/spaces/KB/pages/143163422/Prediction

Tan, K. H., Ortiz-Gallardo, V., & Perrons, R. K. (2016). Using big data to manage safety-related risk in the upstream oil & gas industry: A research agenda. *Energy Exploration & Exploitation, 34*(2), 282-289. 10.1177/0144598716630165

Yang, Z., Chang, Q., Djurdjanovic, D., Ni, J., & Lee, J. (2007). Maintenance priority assignment utilizing on-line production information.(author abstract). *Journal of Manufacturing Science and Engineering, 129*(2), 435. 10.1115/1.2336257

Zemenkova, M., Zemenkov, Y., Pimnev, A., Kurushina, E., & Zemenkova, M. (2016). *System of controlling the reliability of hydraulic machinery in oil and gas facilities*10.1088/1757-899X/127/1/012055